

The University of Chicago

A METHODOLOGICAL STUDY OF THE
LEARNING OF CHEMICAL CONCEPTS
AND OF CERTAIN ABILITIES TO THINK
CRITICALLY IN FRESHMAN CHEMISTRY

A PART OF A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE
SOCIAL SCIENCES IN CANDIDACY FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATION

1944

By
HERBERT A. THELEN

Reprinted from
THE JOURNAL OF EXPERIMENTAL EDUCATION
Vol. 13, No. 1, September 1944

The University of Chicago

A METHODOLOGICAL STUDY OF THE
LEARNING OF CHEMICAL CONCEPTS
AND OF CERTAIN ABILITIES TO THINK
CRITICALLY IN FRESHMAN CHEMISTRY

A PART OF A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE
SOCIAL SCIENCES IN CANDIDACY FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATION

1944

By
HERBERT A. THELEN

Reprinted from
THE JOURNAL OF EXPERIMENTAL EDUCATION
Vol. 13, No. 1, September 1944



A METHODOLOGICAL STUDY OF THE LEARNING OF CHEMICAL CONCEPTS AND OF CERTAIN ABILITIES TO THINK CRITICALLY IN FRESHMAN CHEMISTRY¹

HERBERT A. THELEN
University of Chicago

INTRODUCTION

The educative process provides students with certain opportunities for experiences that are expected to result in desirable development on the part of the learner. The understandings that govern the selection or planning of these opportunities constitute the rationale of the particular teacher. In general, the sources of rationales in education are personal experience, tradition, and authority, with the addition of a large "dash" of expediency. Such rationales are not necessarily "bad", but they are often in conflict with one another, because they are subjective, and they may be scientifically unsound, because their sources are scientifically unsound. A proper theory of education, on the other hand, has its sources in objective experimentation, and it keeps the facts distinct from its assumptions, although it comprises both. Sound educational theory is not "true" in any absolute sense, but its conceptual components are consistent, its salient relationships are demonstrable, and its scope is comprehensive enough to describe and account for all known factors affecting the educative process. The development of sound comprehensive theory and, through it, the improvement of educational practice are the primary aim of research in education.

The link between theory, stated at a general and hence abstract level, and practice, which can exist only at the level of operation, is method consistently implemented with appropriate learning procedures. Within the limits of the experiment described in the present report, the experimental method and procedures were designed to put into practice a broad theory of science education that attempts to comprehend the nature and control of behavior, the nature of learning in science, and the personal-philosophical and

¹ "An Appraisal of Two Methods for Teaching Scientific Thinking in General Chemistry". Unpublished Ph.D. dissertation, Department of Education, University of Chicago, 1944. (Under the direction of Professors Ralph W. Tyler, Karl J. Holzinger, and Stephen M. Corey.)

societal values whose operation is essential to personal and societal security. The control method and procedures put parts of the theory into practice, but their source in uncontrolled experience precludes a close connection with more than fragments of the theory.

PURPOSES

The purposes of this experiment may be stated at the levels of practice, method, and theory:

1. At the level of practice the major purpose is to demonstrate that, by modification of the existing procedures of instruction, a wider range of abilities in the area of scientific thinking can be learned without the "sacrifice" of subject-matter content.

2. At the level of method the major purposes are: (1) to devise a method of teaching chemistry that is consonant with a stated and consistent theory,² and (2) to make concrete suggestions for the improvement of freshman chemistry within the present administrative framework of instruction.

3. At the level of theory the major purposes are: (1) to formulate a general theory of science education (outlined in the original report), and (2) to validate certain theoretical principles by demonstrating that (a) they may be used to predict the results of instruction, and (b) they may be arrived at inductively from reflection upon the learning experiences and the outcomes of teaching.

DELIMITATION

1. The population of the experiment represented about 60 percent of the freshman students enrolled in afternoon sections of general chemistry for engineers, chemistry 114-124, during the academic year, 1941-1942, at Oklahoma Agricultural and Mechanical College, Stillwater, Oklahoma.

2. The methods differences were applied during the four-hour quiz-laboratory sections

² Chapter II of the complete report.

attended once a week by each student. The students were not segregated during two one-hour lectures attended each week.

3. The objectives studied were behaviors that both methods strive to realize. The evaluated abilities may be stated generally as "ability to identify the most appropriate meaning of terms" (knowledge of concepts), and five abilities composed of specific skills believed to be relevant to "scientific thinking".³

4. The subject matter studied was the same in all classes, although its organization was somewhat different under the two methods.

5. The evaluational evidence was obtained from a four-hour objective test administered at the beginning, middle, and end of the year; 31 scores for each student were obtained from each administration of the test. Each score provided a quantitative description of presumably homogeneous kinds of responses to items in the six sections of the criterion test.⁴

6. Other evidence used to describe the students and to assist in some of the interpretations came from a questionnaire made by the experimenter and from Progressive Education Association Test 8.2a, Interest Index; these were administered at the close of the experiment. Also available were scores from the tests of the freshman week battery: A.C.E. Psychological Examination, 1939 (L, Q, and total scores); Cooperative English, Algebra, and Arithmetic Tests, 1939 (total scores); and either the Iowa Chemistry Training Test, 1930, CT-Y or the Iowa Chemistry Aptitude Test, 1930, CA-X (all scores).

INSTRUCTIONAL METHODS

1. The control method was the set of procedures that have been worked out at the College over a period of at least ten years. Two control groups were employed: (1) "semi-micro", making use of small-scale lab-

oratory equipment (for economy) and a laboratory manual that adapted the experiments for use with the midget apparatus; and (2) "macro", making use of the conventional-sized equipment and a laboratory manual to guide the experimentation.

The control groups spent the first hour of the quiz-laboratory meeting in reviewing topics discussed in previous lectures and summarized on a one- or two-page "quiz outline" prepared by the instructor in charge of the course. The remaining three hours were spent in the laboratory following the directions in the manual, recording observations, and writing answers to "interpretative" questions. In the semi-micro groups, an additional feature was the writing of answers to several "preparatory" (orienting) questions before performing the experiment.

In general, it is probably fair to say that the control method was closely representative of at least 80 percent of the instruction now being given in freshman general chemistry.

2. The experimental method made use of the same subject-matter content as the control method, and the experiments under the experimental method involved about the same chemical materials and processes. No laboratory manual was used by the students. Each experiment was written up completely in proper form.

The experimental groups ordinarily participated in a series of activities planned for the following purposes:

- a. Review of lecture material (optional).
- b. Follow-up of preceding week's experiment.
- c. Drill or special activity focused on one of the major objectives of scientific thinking.
- d. Group planning (class or small groups) of experimental work.
- e. Performance of experimental work (pairs, small groups, individuals).

The activities designed for these purposes were quite varied, including such features as silent demonstration, volunteer individual demonstration, black-board drill, student chairmanship, and a number of special mimeographed guide sheets given to the students.

³ This does not mean that these are the only objectives striven for under the two methods. Learning experiences, like other experiences, according to our theory, in either planned or haphazard fashion will affect the interests, abilities, appreciations, attitudes, and other behavior determinants of the students, depending upon their areas of insecurity. A method that promotes critical thinking, for example, at the cost of such outcomes as social sensitivity, creativeness, cooperation, tolerance, self-direction, and aesthetic appreciation, is decidedly a poor method. See also *Science in General Education*, pp. 41-45. New York: D. Appleton-Century Co., 1938.

⁴ Each homogeneous sort of response is assumed to reflect a specific kind or aspect of a learned behavior organization. Such organizations or abilities are in turn objectives of the course. Thus each ability is described from a pattern of several scores.

FEATURES OF THE EXPERIMENTAL METHODS

1. *The class should value the learning objective for each activity.* Such approval is sought through:

a. Periodic class discussions appraising:

1. Knowledge and abilities attained so far in the course.
2. Re-interpretations, as course outcomes, of the importance of various sorts of abilities.
3. Continual reformulations of principles in the light of additional knowledge.
4. Consideration of general programs of study worth pursuing next.

b. A few minutes at the beginning of each activity (or other strategic time) to establish a favorable "set" and to develop recognition of the need for the activity.

c. Incidental identification of real behaviours as planned-for outcomes in specific activities.

2. *The class should thoroughly understand the procedures of each activity.* This understanding is the result of:

a. Planning the activity in group discussion; as facility in planning group action grows in the group, the requisite detail-ness of the planning decreases:

1. At the beginning of the year, all aspects of the plan came from student suggestions, modified by group criticism summarized by students and written in notebooks and on board.
2. At the middle of the year (or whenever deemed reasonable), problems were parceled out to small groups for the entire planning and execution; the problem was concluded in report to and discussion by the class.
3. At the end of year (or whenever deemed reasonable), planning was by individuals with reference to standards and responsibilities set by group.

3. *The class should learn to attack all problems in a sound scientific manner.* This entails:

a. Stating the problem for experimentation. The problem was usually a relationship to be formulated.

b. Listing all factors believed to affect the relationship, and suggesting the nature of the mechanism for each.

c. Evaluating each factor through prediction of the consequences of its variation during the experiment and deciding:

1. Which factor to vary, and which to observe.
2. Which factors should be held constant.
3. Which factors may be assumed to have negligible effect.
4. Which factors are irrelevant.

d. Planning a procedure to enable proper variation, observation, and control of factors.

e. Conducting the experiment, collecting data.

f. Organizing data, and inducing generalizations.

g. Criticizing the experiment, pointing out possible improvements, and estimating magnitude of errors.

h. Applying learned generalizations in new situations.

4. *The class should consciously function as a primary group.* This entails:

a. Deciding as a group when group action is appropriate to achieve desired specific outcomes.

b. Parceling out responsibility for parts of extensive experiments to volunteers, and then integrating and organizing the individual contributions. Personal values consequent upon the experience of reporting to the group, and saving of time for other valued activities (experiences), are motivating factors.

c. Assisting individuals to formulate their ideas, applying criteria of tenability and procedure, and appraising resulting concepts through application of group-determined standards.

d. Applying group censure or encouragement to students whose behavior warrants.

e. Working under student chairmen in activities in which the teacher (as expert) believes chairmen will be effective.

f. Formulating as a group the rules for individual conduct in such matters as laboratory house-cleaning, where the

need for such rules arises from the experience of the group. Delegation of responsibility for the enforcement of such rules to members of the group.

5. *The class should follow up each experiment by analysis* (in group discussion subsequent to individual written reports) of:

- a. Sources of error, due to poor technique, design of apparatus, or order of procedure. Magnitude of the error is estimated on common-sense grounds (of previous experience in and out of class).
- b. Formulation of reasonable generalizations from experimental data.
- c. Suggestion of subsequent appropriate activities.
- d. Identification of cultural areas in which discovered generalization seems operative, and of values implied in its operation and control.
- e. Continual re-examination of previous knowledge, resulting in its reorganization and more precise definition.

Although the comparison of the outcomes under the methods was limited to objectives in the area of scientific thinking, the experimental method was designed to contribute to the realization of many other partially identifiable objectives. The class was regarded as a primary group; emphasis was placed upon acquisition of responsibility for collection of particular data, and for leading discussion by individuals in class. In discussions the "slanting" tended to be toward societal purposes, and philosophical orientation. The possibilities in these areas, it is believed, were scarcely explored. The problem of providing "rich" activities and of guiding participation of students into channels most worthwhile to them emerged as the central problem involved in the use of the method.

The experimental method as here employed was consistent with the practices developed under the leadership of L. F. Foster at University High School, Oakland.

DESIGN OF THE EXPERIMENT

1. *Formation of the samples.*—An effort was made to assign the students to particular quiz-laboratory sections in such a way that there would be approximately congruent distributions of the traits measured during freshman week in each section; presumably, by

this means the largest samples could be obtained. The procedure failed, because the assigned times could not be utilized by many of the students, since many did not complete their enrolments, and because a large number of students enrolled in the course at the conclusion of freshman week. The samples as finally composed represent a population of students whose initial characteristics were between those of students who take one semester and those who take a year of chemistry. There was only about a 60 percent survival, and the "year" students were superior to the semester students. To facilitate comparisons of growth during the two semesters, some of the poorer students were eliminated from the semester samples, and some superior students were eliminated from the year samples. The distributions of initial scores for general accuracy on the six sections of the criterion test were very similar among the three semester and the three year samples, so that gains during the second semester could be inferred from measured gains during the first semester and during the year.

The samples were stratified with respect to per cent of students who had taken high school chemistry, per cent of students taught by each instructor, and roughly by per cent of girls.

2. *Administration of instruction.*—The instruction in the control groups was guided by the professor in charge of the course through the medium of the weekly typed quiz-outlines, and through bi-weekly staff meetings during the first semester. The laboratory instruction was confined to hints, warnings, and mechanics; the quiz-discussion was guided by the instructor as he wished, and ranged from repetition of the regular lectures to occasionally irrelevant discussions. This conformed closely to the usual procedure over the preceding years.

The instruction in the experimental groups was guided by typed, detailed lesson plans devised by the experimenter. There was seldom more than ten minutes of time during the four hours that was not planned for, and in spite of the greater complexity of the activities and the wide differences in understanding by the instructors of the range of objectives, it is probable that there was as much uniformity of overt activity in the experimental groups as in the control. Salient

features of the lesson plans were discussed, usually for about twenty minutes, with each instructor. The experimenter also asked for and received help from the instructors in planning follow-up activities for the experiments, and in deciding when a "change of pace" in the types of activities seemed psychologically appropriate.

3. *Teachers*.—The use of eight different teachers to give instruction during the year may have reduced differences in gains among the methods. The teachers were chiefly young instructors who were sympathetic to the experiment, but whose previous training and thinking had been almost exclusively in chemistry or biological science. It seems reasonable to conclude that any systematic differences between outcomes of the methods were due to the methods. In this case, therefore, it was demonstrated that the experimental method was "practical", since inexperienced teachers with little supervision were able to apply it.

4. *Subject matter*.—An effort was made to treat the same subject matter under both methods, so as to rule out possibly unique interactions with subject matter as the cause of learned differences in behavior. This need for covering a wide range of subject matter, determined without much reference to educational objectives, is one source of discrepancy between the ideal and actual forms of the experimental method.

5. *Collection of evidence*.—The pre-, mid-, and post-test was the criterion battery referred to above. This test was constructed along the lines identified and established by Ralph W. Tyler: general description of the ability to be evaluated, identification of specific behaviors to be taken as evidence for the ability, and description of the ability from a pattern of sub-scores. The validity of such a test at the operational level is thus "built into" the test, since the test and criterion are identical at this level. Other evidence of validity of the test is the agreement between prediction and results of testing.

Data from a questionnaire dealing with hobbies, favorite sports, academic plans, etc.; from PEA Interest Index 8.2a; and from several of the freshman week tests are presented in describing the students fairly completely, so that the type of students to which

the findings are applicable may be known.⁵ These data are also used to describe slight differences among the methods groups; these differences, however, are treated as randomly distributed uncontrolled factors affecting the estimated error variance. In a few cases knowledge of these differences assisted in the interpretation of difference among patterns of scores.

METHOD OF THE EXPERIMENT

The major requisite of the method is that it leads to valid interpretations of observed data. Such interpretations are not "true", but only probable, since human experience has a large number of dimensions and hence many degrees of freedom that may be operated on by "chance". In physical science, validity is obtained through the use of valid procedures; it is assumed that, if every step in the procedure is demonstrably sound, then the results must be sound. The counterpart of this view in education would hold that, if the measuring instruments define the behaviors to be discussed, and if the operational details of the methods are clearly different and practical, then the results should be valid (assuming reasonable control in selection of samples). If, however, the anticipated interpretations of the data are theoretical principles, then there is a much more powerful concept of validity available; this concept arises deductively from the nature of theory. It is generally agreed that sound theory: (a) is consistent with other sound and relevant theory, and (b) may be used to predict results in situations similar to those from which the theory derives. Although any one statistic may agree with theory by "chance" alone, a pattern of anticipations is most unlikely to eventuate through operation of the randomly distributed uncontrolled factors that we call "chance", for, by definition, chance cannot be structured in any anticipated fashion.

In this experiment both concepts of validity are employed. First, the measuring instruments presumably elicit (via paper and pencil) specific behaviors that are observed in the classroom. The methods are operationally described, and it can be shown that the activities of each day are in harmony with

⁵ The intelligence of the students was close to the national average, but was considerably more "quantitative" than "verbal". The interests were preponderantly "practical science". More than half of the students held approximately half-time jobs. Their cultural interests reflected those of midwestern towns (median population of 10,200).

the stated criteria of each method. The samples of students are composed of similar strata, and are shown by analysis of variance and the L-test to be closely representative of the same population. Second, this experiment is concerned with a teaching-learning situation that is typical of those for which the theory of science education is relevant. Therefore, the theory may be used to set up a large number of hypotheses describing the results anticipated through application of the theory. The hypotheses, of course, must be stated in such language that it is possible to assess the degree of agreement between results and prediction.

The hypotheses of this experiment are of several sorts:

1. *Specific hypotheses*.—Prior to the conduct of the experiment, but after the methods had been described, it was possible and appropriate to predict from knowledge of theory that the experimental method should prove superior to the control in increasing status with respect to all but four of thirty-one measures of the criterion test. After the experiment was concluded, but before the tests had been scored, it was possible and appropriate to examine the number of activities in which each of the methods differences had been applied and to estimate the reliability of the anticipated differences. This additional refinement made the hypotheses more precise and hence more easily testable. The range of possible reliabilities was divided arbitrarily into three segments expressed as percent level ranges and chosen arbitrarily:

- a. Reliability of differences falls between 0 and 10 percent levels. Ten of the thirty-one hypotheses predicted this order of magnitude of reliability of differences between experimental and control groups.
- b. Reliability of differences falls between 10 and 30 per cent levels. Seven of the thirty-one hypotheses predicted this order of magnitude of reliability of differences between experimental and control groups.
- c. Reliability of differences falls between 30 and 100 per cent levels. The remaining 14 of the 31 hypotheses predicted this order of magnitude of reliability of differences between experimental and control groups.

These three ranges may be labelled respectively as standing roughly for "significance"; "fairly reliable, significance could probably be achieved by better saturation of methods"; and "unreliable, better saturation of methods differences in subsequent experiments may or may not produce more reliable differences". However, it is not necessary to attach labels to these ranges, since their purpose is mainly to provide more definite criteria for checking.

The reliabilities of the differences between the two control groups were predicted to fall in the range from 30 to 100 percent.

Since pre- and post-scores were obtained, it was deemed appropriate to employ a statistical method that takes account of differences in gains. Since it was anticipated that there would be some correlation between pre- and post-scores (especially in the control groups), it was presumed that part of the difference between pre- and post-scores could be predicted by means of regression alone. Furthermore, the samples were known to be closely similar initially. These considerations make applicable the use of the method of analysis of covariance, which was used.⁶

2. *Interpretative hypotheses*.—During the formulation of the experimental method, a series of objectives, such as "ability to visualize simple experiments from verbal descriptions", "ability to distinguish effects from their causes and from their illustrations", and the like, was identified. While direct evidence was not obtained from the test scores with reference to development of these behaviors, it was believed that careful study and comparisons among the scores, coupled with analyses of students' experimental reports and other work, could shed light on the relative attainment of these presumed abilities. Actually, however, this list was of considerable importance in guiding instruction, but evaluations of these hypotheses were not attempted.

3. *Methodological hypotheses*.—The confirmation of a large number of the specific hypotheses, plus satisfactory subjective evidence that other objectives were attained as

⁶In a few cases in which the correlation between initial and final scores in the population was low, and in which the gains were approximately the same size as the initial differences, this method actually imputed superiority to the inferior group. These anomalies occurred only in the 30 to 100 percent range and are of little consequence, since such differences are not informative. The existence of such occasional anomalies might be obviated by use of separate b's for each group; their existence is not considered to be sufficient justification for the usual practice of assigning an arbitrary level as "significant" and rejecting all differences at a higher percent level.

well or better in the experimental groups as in the control group, could be regarded as some evidence that the experimental method is a "better" method of teaching chemistry. It is probably safe to say that the principles stated above as the features of the experimental method are revealed as leading to more effective instruction than do the features of the control method.

4. *General hypothesis.*—The basic underlying hypothesis of the experiment is that "through employment of instructional activities guided by the experimental method, students will learn as much or more 'subject matter' as under the control method, and will make substantially greater progress in development of desirable specified abilities in the area of 'critical thinking'." This hypothesis is confirmed, although the meaning of "substantially greater" is dependent upon the assumptions one makes about "educational importance" as outlined below.

SUMMARY OF FINDINGS: AGREEMENT BETWEEN SPECIFIC HYPOTHESES AND RESULTS

In the comparisons between the experimental group and each of the control groups, it was found that 13 hypotheses were correct, and that in 9 cases the reliability of the difference in gains was greater than predicted. This means that 71 per cent of the hypotheses were either correct or conservative and, in general, the achievement of the experimental group was slightly better than anticipated. Of the remaining 18 comparisons of experimental group with each of the two control groups, 4 hypotheses called for more reliable differences than found, and the remaining 14 identified the wrong group as superior. Of these cases of wrong identification, 11 concerned reliabilities at the 30–100 per cent level and 3 at the 10 to 30 per cent level.

In the comparisons between the two control groups, 21 hypotheses correctly predicted that the difference in gain would be reliable only between the 30 and 100 per cent levels. Eight of the 9 differences reliable at the 10–30 per cent level favored the semi-micro group, and the one difference reliable at the 0–10 per cent level favored the macro group.

The 31 scores consist of 6 scores for general accuracy (one score for each section of the test), 16 subscores appraising specifically described types of accuracy, scores for over-

cautious and beyond data behavior on three sections, scores representing two other specific errors, and one score for failure to mark the items in a section because of alleged lack of knowledge. The general conclusions from these comparisons are:

1. There were differences in instruction under the two methods.
2. There was correspondence between the nature of the learning experiences and the consequent learnings.
3. The bases for prediction of the extent of correspondence were pragmatically sound, implying that:
 - a. The identification of the experiences relevant to differences in achievement was accurate.
 - b. The estimates of the magnitudes of differences in instruction required to produce reliable differences in outcomes were in the main correct, but tended to be conservative.
 - c. The aspects of theory linking classroom experiences with achievement, in the main, were sound, since predictions were possible.

EDUCATIONAL IMPORTANCE OF THE DIFFERENCES IN ACHIEVEMENT UNDER THE TWO METHODS

Data are presented in the complete report to show that the mean gain of the group making the larger gain in most cases falls well within the range of minus one sigma to plus one sigma of the distributions of gains in the other groups. Furthermore, it is found that the greatest gain of any group on any score is 58 per cent of possible gain, and that the mode is more nearly 25 per cent of the possible gain.

The "educational significance", generally thought of as the amount of difference in achievement per unit of teaching energy expended, is a concept of minor importance in an experiment of this type. Among the reasons for this assertion are:

1. The purpose of this experiment is to describe correspondence between modification of instruction and modification of outcomes. The differences in outcomes are not required to be larger than required for "significance" by the method of covariance and, indeed, there is no advantage in the establishment of

the relationships to finding reliabilities of differences at less than the 1 per cent level.

2. This experiment is not at all concerned with the cost of each difference produced, although it is interested in discovering whether the experimental method is "practical".

3. This investigation has tested only one instance of the experimental method. In any other instance, the method would be the same, but the procedures would probably be different. In other words, the degree of saturation of the methods is not rigorously defined, so that it is not possible to claim that the "educational significance" of differences in achievement can be used to appraise the methods being compared. (However, they would appraise the two series of procedures used in this particular experiment.) The writer's appraisal would be that the control methods, making use of relatively few oft-repeated procedures are nearly saturated, whereas the experimental method as here applied nowhere approaches saturation.

4. Much work would be required to decide whether the learning outcomes in the experiment are primarily convergent or divergent phenomena. (This is outside the scope of this study.) A 10 per cent increase in ability to interpret data, for example, might imply a change in outlook that would increase throughout a lifetime, or it might imply the partial mastery of a particular specific behavior that the student will never use again. The justification for research of this type requires the assumption to be more nearly like the former than the latter evaluation, and there is some indication within the experiment that this assumption is appropriate.

RESULTS AND INTERPRETATIONS

The data from each section of the criterion test, corresponding to each of the six "abilities" studied, have been analyzed in detail. In this report are presented the findings with respect to two theoretical principles, in an attempt to show precisely how these principles operate in the context of freshman chemistry learnings. These principles are concerned with:

1. The relationship between "directness of experience" and the learning of concepts.⁷

⁷ Based on Chapter VI of the complete report.

2. The relationships between experience, verbalization, and "transfer" in the learning of the attitude of "overcaution" with respect to the interpretation of data, the planning of experiments, and the predicting of results of chemical action.⁸

Suggestions are made for modification of the present chemistry curriculum.

I. "DIRECTNESS OF EXPERIENCE" AND THE LEARNING OF CONCEPTS

1. *Nature of the Objective*

Observations of phenomena, descriptions of procedures in the laboratory, methods of working, explanations, and predictions in chemistry are communicated by means of an extensive vocabulary of words signifying concepts whose meanings must be understood. Some of the different sorts of concepts for which words stand may be identified:

a. *Materia*.—"Sodium," "sulfuric acid," and "beaker" are symbols for chemical materials or equipment. They are learned primarily through association of word with object. These words are the simplest to learn, but even these simple words are multi-ordinal; thus a beaker is: (1) a certain shape of glass object, (2) a container for solids and solutions, and (3) a reaction vessel. Further, a beaker may be any one of a number of sizes, and the student must learn to qualify the word "beaker" with such phrases as "250-milliliter." The multiordinality of the terms for compounds is a very common source of difficulty. "Sodium sulfate" may be used to designate the contents of the fourth bottle on the third shelf; it may be regarded as the symbol for a certain white crystalline solid; or it may refer to a complex of properties of a compound, as described below.

b. *Summaries of properties and processes*.—The student of chemistry must learn to associate observed properties with names. This is one sort of ability commonly included under the general heading of "ability to apply principles." Around the term, "sodium sulfate," the student needs to learn to marshal a large cluster of concepts: (1) this compound is a sodium salt, and hence is soluble; (2) it contains sulfate ion, and hence will give a precipitate with acidified barium chloride solution; (3) it may be produced by the action of sulfuric acid on sodium hydroxide;

⁸ Based on Chapter XI of the complete report.

(4) 1 gram-molecular weight of this compound will raise the boiling point of water by 1.56 degrees C, and will lower the freezing point by 5.58 degrees; (5) a water solution of this compound will conduct electricity; and so on, through a long list of behaviors.

The summary of properties may include properties of a single specific compound, or may represent an abstraction, or class-name. Thus "acid" symbolizes a group of commonly observed behaviors, such as turning blue litmus red, catalyzing esterification reactions, and producing hydrogen by action on active metals. "Synthesis" is an abstraction of a common-element in a great many reactions and processes, and "distillation" calls to mind a set-up of apparatus and a sequence of operations. It is probable that these summary words are best learned inductively, for one may experience operationally the properties the words abstract. The sound, "acid," when learned thoroughly, should "call to mind" a sour taste, a glowing light bulb, a warmed test-tube, and/or a stream of gas bubbles, depending upon the gestalt in which the symbol appears.

c. *Theoretical concepts*.—There are a few substances and a great many processes that came into chemistry as a result of the quest for mechanisms to explain and predict chemical behavior. These theoretical notions themselves are not available to sensory perception, but are the result of comparisons and analyses among summaries of observations. "Oxidation," for example, is a broad theoretical concept of increase of valence. The mechanism associated with the effect is that of losing electrons ("electrons" is another theoretical concept). The observation that some substances combine with oxygen is an illustration of the oxidation process, and, at an elementary level, oxidation is sometimes defined as "combination with oxygen". When most usefully understood, these concepts are insights, products of reflection, that may be applied to a wide range of seemingly-different phenomena. It is likely that theoretical concepts are usually learned verbally from lectures, discussion, and text, but they are non-functional, unless the student has had considerable experience in applying them to problems set in the laboratory or described in textbooks and workbooks.

d. *Concepts for convenience*.—"Equivalent weight" and "absolute temperature" are con-

cepts that have theoretical significance, but the student's chief need for them is as a convenience in solving problems efficiently. Students often try to solve problems in acidimetry or kinetic-molecular theory by cumbersome methods involving substitution of less appropriate notions; although the answer may eventually turn out the same, the process is unnecessarily wearisome. In such cases the difficulty is usually a failure in ability to state the precise differences between "equivalent weight" and "gram-molecular weight," and between "absolute temperature" and "degrees centigrade." In general, there is need to know from where the term comes and why it is used.

It is clear that the precise learning of concepts requires not only the application of criteria for generality and specificity of terms, but also appropriate understanding of their operational, relational, and theoretical associations.

2. *Teaching of the Objective*

In helping students learn chemical concepts, the teacher may select from among three courses of action: (1) he may make no special provision of activities designed to lead to desirable habits, skills, and critical abilities in the use of chemical terms; (2) he may seize such opportunities as arise during the class and laboratory work for pointing out generalizations about the use of terms or for making observations about their etymology or definition; or (3) he may carefully plan activities to promote in students various generalized behaviors that his analysis has convinced him are important to the objective.

In this experiment, the teaching of the objective followed plan 1 under the control method as employed in the semimicro and macro classes; this plan involves the assumption that the ability to use terms correctly is an automatic concomitant of the learning of "chemistry." Aside from the inclusion in the laboratory manual of a few questions dealing with definition, little was done that may be expected to contribute directly to the objective. Under the experimental method, the teaching followed plan 2, in that the knowledge of terminology was an objective concomitant to other objectives. A fairer statement might be that the basic assumption in the experimental methodology is that the experiences of the students may be planned

to satisfy simultaneously a variety of objectives, including knowledge of terminology. Accordingly, the devices used to develop knowledge of terminology are casual and usually are merely an aspect of the far richer class activity. Among the devices employed were:

- a. Definitions of terms formulated through class discussion and written on the board.
- b. Application of criteria of definition to the criticism and evaluation of definitions suggested in class.
- c. Discussion of the etymology of such words as "anhydrous."
- d. Recognition of the different purposes and levels of definition, and selection of the most appropriate definition. Thus, "reduction" may mean "loss of oxygen," "decrease in valence," or "gain of electrons."
- e. Provision of opportunity during discussion for students to receive aid from teacher and class in making their meanings clear.
- f. Emphasis upon conciseness and clarity in the individual written laboratory reports. These reports were carefully read and marked by the instructor, and the most common errors and misconceptions were subsequently discussed in class.
- g. Formulation of concepts only after experiences designed to make the concepts meaningful.
- h. Placement of emphasis upon obtaining meaning from the book. Although this too often took the form of exhortation, two special activities (reading the book line-by-line and locating definitions and arguments in the book) were employed.

3. *Appraisal of the Objective*

a. *The objective of desirable changes in behaviors.*—Reflection upon the observed difficulties of students in communicating knowledge suggested that the general behavior which might be used to appraise the objective is the distinguishing of the "best" statement of the meaning of a term from partially-correct statements that students often confuse for the "best" meaning.⁹ Casting the situa-

⁹ This decision includes the assumption that ability to compare given statements correlates highly with ability to volunteer correct statements. There is some evidence in certain kinds of exercises that this is true. The interpretations in this section are confined to the objective as defined by the measuring instrument.

tions in multiple-choice items makes possible the appraisal of ability to make distinctions of any desired degree of delicacy. In the test used, a coarse distinction is the recognition that a "factor" is "anything which may affect the result of an experiment" rather than a "person who seeks facts." A fine distinction involves such things as identifying "weight" as "a force" rather than as "a mass"; the common confusion arises from the fact that "weight" and "mass" are used practically as synonyms in chemistry. Not only may distinctions be delicate or coarse, but they may vary as to their nature, and a test that is to appraise the objective adequately must include situations involving the more important sorts of distinctions. The test contains erroneous statements of the following types:

1. Gross misconception: experimental error as "an experiment performed by mistake," and the use of valence to decide whether an element is "deeply colored or transparent," are examples of misconceptions. The choices involving misconceptions provided an opportunity to suggest non-technical responses: it was expected that the initial frequency of marking misconceptions would be about that expected by chance.

2. Too general statement: "amount" is a too general definition of "volume," a very plausible definition which is too general, because it does not control temperature, is that "air pressure" is a "measure of the number of molecules of air in a given volume." Too general statements embrace too wide a range of occurrences, fail to control important factors, or fail to be sufficiently specific to distinguish between phenomena of the same class.

3. Too specific statement: calling an "atom" "the smallest part of a metal that has the properties of the metal"; restricting "evaporation" to situations involving "high temperatures"; and limiting "synthesis" to "forming a precipitate" are examples of defining terms too narrowly or too specifically.

4. Statement of cause or mechanism for effect: "diffusion" as "the result of molecular impact" and "experimental error" as "due to incorrect way of reading instruments" are examples in this category. Prevalence of this sort of error usually suggests confusion of phenomena with their theoretical explanations or with relevant observations.

5. Statement of illustration for effect: "air pressure" may be illustrated incorrectly as "the weight of air in a tire," "filtration" is illustrated in "separating water from mud." The use of illustrations usually suggests failure to generalize an experience to the most useful degree of abstraction.

6. Statement involving misunderstanding of scientific method: since understanding of scientific method was a major objective in the experiment, it was desired to provide opportunity for the assertion of confusions about the terminology of scientific method. These confusions may include some of the sorts of errors listed above. Examples in this test include: a "fact" is something which "can be proved," and "an experiment" is performed to "prove the law of cause and effect."

Terms may be improperly defined, if the student is confused in any of the ways indicated above. In addition, there are different "levels" of definition; a term may be defined operationally or in theoretical language. Three classifications were used in building the test:

1. Definition through application or function: a "catalyst" may be defined as "something which changes the speed of a chemical reaction," and an "electrolyte" is "a solution which conducts electricity." In these cases, the term acquires meaning through the useful characteristics of the phenomenon represented.

2. Literal definition: words like "hypo-," "anhydrous," and "equilibrium" may be defined on the basis of a knowledge of words rather than of phenomena. Such definitions are thought of as "literal definitions."

3. Technical definition: a highly theoretical statement about a "non-metal" is that "when reacting with other substances, it tends to take on electrons," or that "temperature is a measure of the amount of kinetic energy of the molecules in a body." In general, technical definitions are relational in that the term is defined through its relations with other terms.

The foregoing analysis of behaviors served a useful function in the construction of the test, for it made possible the inclusion of responses that could sample a wide range of possible behaviors. The analysis was also useful in the conduct of class discussions, since the experience of making the analysis left the teacher somewhat more sensitive to specific sorts of difficulties that impair communica-

tion. There are, however, many commonly-held mistaken concepts that represent error of fact or principle, and they are not classifiable under the foregoing scheme, but they are patently involved in communication and, therefore, belong in this section of the test. While the foregoing analysis is suggestive, nonetheless the first criterion to be applied to every item is that it is valid, in the sense that the answers from which the student selects are the sort of things students ordinarily would volunteer. Therefore, this analysis does not define a list of specific objectives for which specific procedures are planned, but it does supply a partial critique for use in teaching the general objective.

b. *The objective of informational content.*—The behaviors listed above can be exhibited only with respect to a situation—a specific pattern of stimuli that call for relevant behavior. The situation set in each test item is defined by the particular connotation of the particular term being tested. In general, the connotation being tested is either the most important one from the standpoint of further learning of chemistry or else it is believed to be the most useful in discriminating between levels of achievement of the objective. (The use of the most discriminating connotation involves the assumption that the behavior it calls forth will correlate highly with the behavior elicited with respect to the most important connotation.¹⁰)

4. *Preliminary Results and Assumptions in Treating and Classifying Data*

One may postulate that, besides the factors controlled by the experimental procedures, there are two sorts of factors affecting the learning of the concepts tested:

1. Differences attributable to the nature of the concepts themselves and to the availability of the concepts through various general learning procedures, such as laboratory, discussion, and lecture.

2. Differences due to the method of supervising and organizing the learning activities in the three methods groups.

Inspection of the data shows that the learning of the concepts tested in Part I of the test was quite similar under the two methods of

¹⁰ A discussion of the procedures used to select representative terms and "reasonable" responses is omitted from this report.

It is interesting to note that about 75 percent of the vocabulary is developed in the first semester, although the "meaning" of most of the terms changes throughout the year.

instruction. On the other hand, great differences in initial scores and gains during the semester and year are noticed from concept to concept, and this suggests that a fruitful analysis of the data would be one which first identified and rationalized the gross differences among concepts and then studied the differences between methods groups as an additional factor.

The first step in the treatment of these data consisted in classifying the results of the testing and in identifying groups of concepts whose learning appeared to be based on similar experiences and whose nature appeared to be much the same. Before proceeding to a study of each category, it is desirable to present evidence that the categories named in the stub of the tables are clearly distinguishable. The outcome of this procedure is shown in Table I.

Table II presents the gain in per cent of the group making the correct response under the experimental, semimicro, and macro methods. The classification of concepts is the same as in Table I.

Tables I and II show that the groups of items classified as indicated behave very differently on the test. Initial means in the categories range from 10 per cent to 76 per cent mastery for the entire populations. Gains range from -7 per cent to 50 per cent in the methods groups. Little relationship is seen to exist between initial and final mean scores among the various categories, thus suggesting that the categories are independent.

The classification: general statement.—The concepts contained in categories A through D may be said to have been "taught." They were presented in lectures attended by the students unsegregated in methods groups. Most of the concepts were discussed or used in discussions of problems in the quiz-laboratory sections, which were segregated into classes under the experimental, semimicro, and macro methods of instruction. In these methods groups, assimilative experiences were directed. Textbook chapters (the same in all classes) were assigned and sometimes discussed; more-or-less assistance was given students working on problems assigned dur-

TABLE I

LEARNING OF THE CATEGORIES OF CONCEPTS BY THE ENTIRE POPULATION OF THE EXPERIMENT

Average Per Cent of Possible Score (expressed as per cent of students marking items correctly)

Classification of Concepts	Number of Concepts	First Semester			Year		
		Initial	Final	Gain	Initial	Final	Gain
A Experimental.....	4	45	75	30	40	88	48
B Functional.....	7	48	65	17	47	71	24
C Theoretical.....	6	10	25	15	10	32	22
D Common Usages....	7	76	85	9	75	86	11
E Methods Implications.....	4	21	24	3	21	28	7
F Not Taught.....	2	24	26	2	29	24	-5
Total, average....	30	41	55	14	41	60	20

TABLE II

GAIN IN PER CENT OF STUDENTS IN EXPERIMENTAL SEMIMICRO AND MACRO GROUPS MARKING ITEMS CORRECTLY

Gain in Per cent of Group

Classification	Number	First Semester			Year		
		X	SM	M	X	SM	M
A Experimental.....	4	41	32	24	47	47	50
B Functional.....	7	21	14	17	34	17	22
C Theoretical.....	6	19	13	13	25	22	18
D Common Usages....	7	10	9	6	16	10	9
E Methods Implications.....	4	6	3	2	6	5	9
F Not Taught.....	2	2	-1	4	-7	0	-6
Total, average....	30	17	13	12	23	18	18

ing lectures; the laboratory experiments were introduced, carried out, and followed up; and any special assignments were made. In addition, various amounts of remedial help were given to students by their instructors outside of class hours.

On the basis of these conditions, certain assumptions appear to be justified:

1. Any differences in outcomes in the various methods groups cannot be attributed to lecture experiences, since the groups were not segregated by methods in the lecture sections.

2. Although the effectiveness of the lecture presentation was probably not the same for the various concepts taught, it is likely that the assimilative experiences in the quiz-laboratory sections determined to a large degree the amount of learning that took place. If this assumption is correct, it should be possible to classify the items according to the general nature of the assimilative experiences and to find variation in outcomes among the categories attributable to differences in the nature of these assimilative experiences.

3. Although the general assimilative experiences (laboratory, as distinguished from discussion, problem work, etc.) may be the same under the various methods, the effectiveness of the experiences may be quite different due to differences in organization, motivation, follow-up, use of review, and the like. Such differences would be revealed mostly in the learning of individual concepts, although general superiority of certain kinds of procedures may be revealed from category to category.

5. Brief Characterization of the Categories¹¹

Category A, "experimental".—The concepts classified in this category are characterized as follows:

1. Part or all of a laboratory experiment was experienced by the students for the purpose of learning the concept.
2. The experiment was direct and valid.
3. The concept may be obtained from the percept by simple abstraction.
4. The statement of the concept is couched in simple, relatively non-technical language.

¹¹ The original treatment of this section includes the following information about each of the thirty concepts tested: the concepts thus categorized, information about their learning as gleaned from analysis of right and wrong responses, information about the experiences likely to have resulted in the learning, and whatever evaluations of the effectiveness of the methods seem reasonable.

Summary, "experimental" category.—The data on the learning of the four concepts most directly illustrated through individual experiments are striking. In this category fall three of the five items of the entire test on which the gain of all three groups was significant at the one per cent level over the year's work. The learning of the fourth concept in this group was significant for all three groups at the five per cent level; only five additional items in the test revealed gains at this level of significance for all three groups. These findings lend support to the notion that laboratory experiences from which concepts may be obtained through simple verbalization are most effective in promoting learning.

In the experimental methods group, the procedures were designed to provide a closer connection between the laboratory experiences and the subsequent formation of concepts from these experiences. It is reasonable, therefore, to expect the experimental group to gain more significantly than the other groups. Table II shows that the superiority was marked over the first semester, but that all groups had reached about the same degree of proficiency with the "experimental" concepts by the end of the year. Consideration of the nature of the learning experiences provided for the formation of the four concepts shows that two mechanisms were operative in leading to the superiority of the experimental sections: (1) in the case of "amorphous" and "electrolyte," the concepts were studied experimentally by the X groups, but not by the semimicro and macro groups during the first semester; (2) in the case of "anhydrous" the experiment was more effective in the experimental groups than in the other groups.

Category B, "functional".—The concepts classified in this category are characterized as follows:

1. The concept is necessary for the description or simple explanation of phenomena.
2. The concepts are stated in simple language and probably involve little reorganization of experience for their recognition.
3. The concepts have been illustrated in aspects of laboratory experiments, but it is likely that the relationship between concept and experiment has not been made clear.

These criteria distinguish between the "experimental" concepts and the "functional" concepts in terms of the closeness of their association with particular experiments. They are designated as "functional," because they are among the handy conceptual tools needed for chemical communication; a person who cannot use these concepts may be regarded as chemically illiterate.

Summary, "functional" category.—Table I indicates that the learning of the "functional" concepts, while greater than in some other categories, was much less than the learning of the concepts directly demonstrated in experiments. Table II shows that the experimental groups were the most successful in the learning of these concepts, and analyses of the learning of each concept suggest that the reason is the closer relationship between the concepts and laboratory experiences.

The learning of the concept of "synthesis" illustrates the possibility of a simple notion remaining vague and not becoming associated with any definite sort of reaction, even though the students certainly worked with such reactions. The learning of the meaning of "one equivalent" is even more striking, for in this case the purpose of one experiment was to clarify the concept. The fact that the definition required in the test was at a more theoretical level than the definition presumably learned from the experiment may account for the poor learning; on the other hand, failure to be able to reformulate the outcome of the experiment in terms of the test item shows how superficial the learning from the experiment actually was.

In general, the analysis of the learning of this category suggests very clearly that it is possible to focus experiments in such a way that the learning of important common concepts may be far more effective. Another way of stating the conclusion would be that the experiments, guided primarily through the laboratory manuals, were less efficient as aids to learning than were similar experiments, developed and followed up in class discussion.

Category C, "theoretical".—The concepts classified in this category are characterized as follows:

1. They represent "new" learning, in that the initial scores on the items were considerably lower than would be expected

by chance. For practical purposes, it may be said that the students had no initial knowledge of these concepts.

2. The concepts are theoretical and tend to be non-operational. They cannot be demonstrated experimentally in any direct fashion, nor do they acquire any particular significance through reflection on common experience (with one exception).
3. Little use of the concepts was required in the course. They were not mastered to the point of usefulness in prediction and explanation, although they were used for these purposes from time to time.

Summary, "theoretical" category.—Inspection of Table I suggests that the learning of the theoretical items is about as great as the learning of the functional items, in spite of the more tenuous relationship of the concepts to concrete experience. On the other hand, the gain for the year on the functional items was 24 out of a possible 53 per cent, or 45 per cent, whereas the gain on the theoretical items by the same method of calculation was but 24 per cent of the possible gain. The apparent superiority of the experimental group is small. Since the theory was related to laboratory work somewhat more in the experimental group, this difference is in line with the hypothesis and is thus more likely to be "real" than would be suggested through application of statistical techniques alone.

Analysis of the interpretations of the learning of the items in this category leads to certain generalizations that seem reasonable in the light of performance:

1. In five of the six items in this category, a majority of the students marked a particular wrong response at the end of the year. In three of these cases the response was actually incorrect; in the other two cases it was partially correct.
2. In general, the most popular response tends to be more operational than the correct response.
3. The learning of the theoretical responses is specific and superficial and useless, rather than profound, insightful, and functional.
4. Learning of theory is dependent upon its use in interpreting concrete experiences.

In general, the learning of the theoretical items is thoroughly unsatisfactory. There is no doubt that theory is difficult to teach effectively, because it involves either: (1) a very broad experimental base which is hard to provide and difficult to interpret; or (2) a great deal of insight, so that relationships may be deduced from definitions. As in all curriculum construction, two questions are pointedly raised: (1) What theory is important (and for what reason)? and (2) How can it be efficiently taught?

Category D, "common usages".—The concepts classified in this category are characterized as follows:

1. The concepts are known initially to a great majority of the students, and are fairly common in everyday communication of ideas, both in life and in chemistry.
2. The meaning of the concepts is little changed by their use in chemical contexts.

The high initial accuracy with these concepts means that a gain significant at the 5 per cent level would require that nearly every student mark the correct statement. The final average per cent mastery for the semester and year groups respectively would have to be 91 per cent and 93 per cent respectively for significance at the 5 per cent level. It is clear that such items are not likely to be discriminating among methods. Since the concepts are so generally understood, no attempt will be made to cite specific learning experiences, except as required for interpretation of results.

Summary, category D, "common usages".—Table I indicates that the terms were known initially to three-fourths of the group. Analysis of the responses shows that the remaining quarter tended to mark relevant responses, suggesting that all the group "had an idea" about the concepts. The gains of 9 per cent the first semester and 11 per cent during the year amount to 37 per cent and 44 per cent respectively of the possible gains. Table II reveals that the gains in the experimental group appear to be slightly higher than those of the other group.

In the case of two of the items, it is suggested that a correct but simple definition whose relevance is not obvious may tend to be eschewed for a more relevant-seeming technical statement. In general, the learning

appears to have progressed during the year. Special learning activities are unnecessary for these concepts, because the student has already had sufficient experience with them to render the incidental contacts in chemistry meaningful.

Category E, "methods implications".—The concepts classified in this category are characterized as follows:

1. The concepts are insights probably developed through reflection upon methods of science.
2. The concepts were not "taught" as verbal statements; the terms were used, however, in context on many occasions.

Summary, category E, "methods implications".—Table I shows that the initial mastery of the concepts in this category is about what one would expect from guessing alone. There appears to be a slight gain during both semesters, but it may not be reliable. There were small but significant shifts in individual items, but they involved few students. In general, one would conclude that the concepts have not been learned, that working with chemistry for a year by no means assures the understanding of even the simplest implications about method. Since the "correct" responses to these four items represent simple concepts, and since the other responses are fairly obviously not tenable, it may be safe to generalize as follows: If understanding of scientific method is an objective of the course, then activities must be especially planned for this objective.

Category F, "not taught".—The concepts classified in this category may be characterized as follows:

1. The concept was not presented in the course.
2. The concept is not a reasonable inference from the experiences of the course.

Summary, "not taught" category.—Table I indicates that the accuracy in this category is about what one would expect from the operation of chance alone. The results show slight shifts, particularly toward "mass" as a definition of "weight." The terms are used synonymously by most chemists, since weight is a measure of mass, although weight is actually a force. The obviously irrelevant responses to weight have not increased, sug-

gesting that there has been some orientation with respect to "weight" during the year.

In general, the performance on the two items of this category suggests the conclusion that concepts not taught will not be learned—certainly a reasonable notion.

6. Summary and Interpretation of Results

1. Comparative effectiveness of various types of learning experience.

a. The items of this part of the test have been categorized under six headings. Average scores on the items in the six categories differ in initial accuracy, final accuracy, and gain during semester and year. The methods of teaching of the items are homogeneous within each category and different between categories, so that in the absence of any other systematic differences between categories it is reasonable to suppose that differences in learning are associated with the differences in general procedures of teaching.

b. A clear picture of the differences in gain between the categories is revealed in Table III, in which the average gain for the population as a whole is compared with the possible gain defined as the difference between the initial average score and 100 per cent.

The preceding analyses suggest that the gains in Table III may be considered as due to the interaction of two factors: initial knowledge (presented in Table I), and nature of the learning experiences (presented in discussion). The nature of the interaction is clearly revealed by comparisons among the categories, thus:

(1) The initial scores in categories E and F were approximately the values expected from the operation of random guessing. The items were not taught, i.e., although closely related concepts were used in class discussion,

the specific aspects called for in the items could be learned only as insights developed by the students without the aid of specific experiences planned for the purpose. The gain is seen to be negligible.

(2) The three categories A, B, and C may be thought of as establishing a continuum with regard to directness of experience. Concepts in the experimental category A were learned through direct laboratory experience; category B, functional, contains concepts whose learning resulted from daily use in describing or explaining experimental results, i.e., concepts that could be associated with several experiences in the laboratory; category C, theoretical, contains concepts whose learning must be almost entirely the result of vicarious experiences. Table III reveals that the differences in learning are spectacular (particularly for the year). To support the generalization that the more direct the experience the higher the gain, one would have to take account of differences in initial scores, for the magnitude of the gains may be related to the initial scores in one of several different ways:

- (a) If the initial score is high, there is less opportunity for gain and less gain is anticipated.
- (b) If the initial score is high, the student has a sufficiently good conceptual framework that additional learning is easy and a large gain is anticipated.
- (c) Within fairly wide limits, the magnitude of the initial score cannot be used to predict the magnitude of the gain.

Comparisons among the categories of initial scores and gains provide some evidence as to which of the three points of view is most appropriate. In two of the categories (A and B) the initial scores are practically the same

TABLE III
PER CENT THAT AVERAGE GAIN IS OF POSSIBLE GAIN

Classification of Items	Per Cent of Possible Gain		
	Number	First Semester	Year
A Experimental	4	55	80
B Functional	7	33	45
C Theoretical	6	17	24
D Common Usages	7	38	44
E Methods Implications	4	4	9
F Not Taught	2	3	—7
	6	6	6

and the generalization about directness of experience is supported. In category C, the initial scores are very low, the learning experiences are indirect, and the gain is low. The smallness of the gain again accords with the generalization that direct experiences are most effective, and it also agrees with the second assumption about the role of initial understanding, to the effect that little learning is anticipated if the student has no initial conceptual framework. There is no way of telling from these data which factor (initial score, or directness of experience) is primarily responsible.

(3) The gains in categories B and D are very similar, and the learning experiences were of the same order, i.e., in both categories the learning resulted from the need for using the concepts for purposes of describing chemical situations. The initial scores, however, are quite different. This tends to support the third position with respect to magnitude of initial scores, that within fairly wide limits, the magnitude of the initial scores cannot be used to predict the magnitude of the gains. Application of this notion to the comparisons among categories A, B, and C would rule out the effect of initial knowledge as a factor in accounting for the low gain in the "theoretical" category and would leave directness of experience as the primary factor in learning.

c. The foregoing analysis has led to the conclusion that the more direct the experience the greater will be the learning, regardless of the magnitude of the initial knowledge (within limits not specified). At this point it may be well to evaluate the extent to which the conclusion is justified from the experiment. Generally speaking, two considerations enter at this point. First, it must be shown that the items in a given category fairly represent that category. None of the categories contains more than seven items; the seven specific behaviors elicited in such a category would seem to be insufficient to serve as a basis for sweeping generalizations about behaviors. On the other hand, if it can be shown that in each of the categories the sampling is excellent—that each specific item epitomizes its category—then the generalization is much more tenable. Detailed analyses of the results from each item were necessitated by this consideration and represent an effort to present the evidence needed to decide whether the categories are well established in this experiment.

The particular issues in this experiment are: (1) Has it been shown that the learning experiences of all the items in a given category are similar (and different from the learning experiences of the items in other categories)? and (2) Have the categories been correctly characterized in general terms? In the writer's opinion the discussions provide affirmation on both issues.

The second demonstration has to do with the reliability of the experiment. If it is agreed that the categories exist and have been appropriately described, then the next question is: How "real" are the observed differences in learning in the categories, i.e., to what extent can the differences be accounted for by fluctuation of uncontrolled randomly-operating factors? Since the categories are assumed to control factors relating to instructional procedures, the major remaining possible source of uncontrolled factors is the individual learner. Since there are 144 learners in the semester group and 84 in the year group, it is highly unlikely that the same chance fluctuations would operate in the same way with any noteworthy fraction of the population; therefore, the differences in gains could be due only to systematic differences in learning procedures. The writer believes the conclusion that directness of experience is the major factor in the learning is justified from this experiment. The fact that this generalization is generally accepted in educational psychology increases the validity of the conclusion.

2. *Comparative effectiveness of the three methods of teaching.*—The preceding analyses suggest that the most effective learning of concepts comes through direct experience in experimental situations designed to facilitate the learning of the concepts. Next in effectiveness appear to be experiences eliciting the use of the concepts for communication of description of experimental procedures, observations, and results in a wide range of situations. Less effective is the vicarious verbal experience with theoretical concepts that are not connected functionally to any operations by the learner. Least effective is poorly motivated learning through induction from experience of concepts that are not held as objectives by the students.

The rationales and the teaching procedures of the various methods as described above suggest two bases for prediction that the experimental groups would demonstrate supe-

rior gains: (1) On the assumption that the foregoing hierarchy of effectiveness exists, the learning experiences were planned at a more operational level in the experimental groups; and (2) when the experiences were at the same level in the hierarchy they were expected to be more effective in the experimental group, because the objectives of the day's work were more clearly identified and striven toward. The data are inadequate to indicate what portion of any observed superiority may be due to either factor alone.

a. One procedure for comparing the outcomes under the three instructional methods would be to tabulate the number of gains (at some arbitrarily chosen level of significance) in correct responses made by the groups. (The results of the procedure will be found in the original report.)

b. A more statistically adequate procedure for comparing the gains under the three methods of instruction is through application of analysis of covariance, since the comparison is then not confined to gains at the 5 per cent level or better. Consideration of all the data might be expected to produce more reliable differences among the methods. Table IV compares the average gains in general accuracy (right responses) among the three methods.

cated in Table IV. It appears that the hypotheses about achievement on Part I of the criterion test have been confirmed, within the statistical and other limitations of the experiment.

3. Summary of Conclusions

a. The rationale of the experimental method includes the initial assumption that learning is most effective when it comes through direct sensory experience of the learner. This assumption was regarded as an hypothesis and tested through analysis of the learning of chemical concepts as measured in Part I of the criterion test. A hierarchy of experiences arranged in order of effectiveness was identified, and the existence of this hierarchy appears to support the initial assumption and thus to add validity to the rationale.

b. A number of interpretations of the course of the learning of various concepts were attempted. These interpretations are guesses arrived at inductively from detailed study of single items, but they may be useful as hypotheses for further investigation. Among these possible hypotheses are:

(1) In proceeding from wrong to right responses with regard to a concept, the students are likely to pass through a stage in which the wrong response is avoided and relevant

TABLE IV

LEVEL OF SIGNIFICANCE OF DIFFERENCES IN GAINS AMONG METHODS, GENERAL ACCURACY, PART I, BY ANALYSIS OF COVARIANCE

(greater)	Experimental vs Semimicro	Experimental vs Macro	SemiMicro vs Macro
(lesser)			
First Semester	14%	2.3%	43%
Year	1.5%	10%	—40%

Table IV shows that the reliability of the differences in gains between the experimental groups and the other groups is quite marked, whereas the superiority of the semimicro groups over the macro groups is unreliable. The differences are more reliable than the hypotheses anticipated: according to the specific hypotheses, the experimental group was expected to be superior to the other groups at a level of significance between 10 and 30 per cent; actually, in two out of four cases, the reliability is considerably greater than this. The specific hypotheses show that no difference between the gains in the semimicro and macro groups was anticipated, and this agrees well with the 40 per cent level indi-

but incorrect responses are accepted. In general, "relevant" in this connection means that the response refers to a class of substances or a process with which the student has had experience at the time of exposure to the concept. It is possible to accept and subsequently reject more than one relevant response, depending upon the nature of the experiences with the concept that one has at different times. During this period it seems likely that the concept is purely connotative; it is connected with isolated experiences and there is no apparent success (probably no effort) to rationalize the concept into any consistent chemical scheme. This period may last up to a year at least (this study has no

evidence beyond a year). Selection of the right response may indicate that: (1) the student has rationalized the concept; (2) the student has learned it verbally; or (3) the student has finally "discovered" it by abstraction from his various experiences with it.

(2) Opportunity for effective learning of a particular concept does not necessarily mean that a closely related concept will be learned. Thus one definition of a term may be learned through experiment and another definition of the same term, at a slightly more theoretical level, may not be understood any better than formerly. There seems little doubt that integration of relationships by the students is far from automatic.

(3) When students learn a concept first at the operational level, and later at the theoretical level, they appear to be confused. It is suggested that an important adjunct to teaching of this sort of concept would be the teaching of the nature of definition and its various levels, and the classification of definitions according to their level in the operation-theory continuum.

(4) Mere theoretical study of a concept is relatively ineffective in correcting a popular misconception.

(5) There may be a tendency for students to abandon popular but correct concepts for more technical relevant but incorrect responses. Laboratory experiments may be able to raise doubts in the minds of the students about points that were once fairly obvious to them.

(6) The difference in gain in general accuracy of the experimental groups over the other groups is statistically significant between the 1.5 and 14 per cent levels. There is no reliable difference in the gains of the semimicro and macro groups. It seems reasonable to conclude that the experimental method of teaching, as outlined above, is more effective than the methods in the semimicro and macro groups with respect to the objectives measured in Part I of the criterion test.

II. THE RELATIONSHIP BETWEEN EXPERIENCE AND VERBALIZATION IN THE LEARNING OF THE ATTITUDE OF OVERCAUTION

In this experiment, it has been useful to recognize that there are two kinds of methods differences: (1) the general or summarized differences between groups of students taught

by different methods, and (2) the specific differences within any particular group. These features are illustrated in the treatment of data in the preceding section. The first kind of difference is given by the statement that the reliability of the difference in gains between semimicro and experimental groups was at the 1.5 per cent level. The second kind of difference is given by the statement that within the entire population of the experiment, the gain in learning of concepts closely related to experiments was almost twice the gain on concepts used mainly for communication of chemical information and procedures.

By making one reasonable assumption, that the abilities to be discussed now were not "taught" in the control groups, it is possible to regard differences in gain between experimental and control groups as a function solely of the "teaching" under the experimental method. Since the activities employed in the experimental method are known in detail, and since these activities were guided to the best of the writer's ability by the stated philosophy, it should now be possible to compare differences in the learning of various abilities between the experimental and control groups to check aspects of the theory, as applied in the instance of this experiment. If, in addition, it can be shown that the conditions of learning each ability were typical of a population of similar sets of conditions, then one is in a position to generalize the results of this experiment to the level of hypotheses about the factors affecting the learning abilities in general. While such generalizations would not have been "proved" in this experiment, their agreement with principles stated would tend to increase the validity of the principles. It may be of interest to note, incidentally, that many of these principles were used in making the predictions incorporated in the specific hypotheses, and it would be most unlikely that an induction of principle from observed differences would not agree closely with principles that accurately predicted the differences. The induction from observed differences is more elegant, in that it starts from quantitatively described observation, and it is probable that the predictions themselves were based upon indirect application of recollected inductions from analagous cases.

1. *Learning of abilities: amount of experience.*—Examination of the reliability of the differences in gain on general accuracy with each of the abilities studied provides opportunity for gross estimation of the effect of amount of experience in learning of abilities. While other factors to be discussed below modify the picture considerably, the general outline at least is clear.

The students had the greatest amount of experience (as judged by time so spent) in planning experiments; next came interpretation of data. For these two abilities, the reliabilities range from levels below 1 per cent to 1.5 per cent. Prediction of chemical occurrences was a component of other activities, and the reliability of the difference in gains is quite different in the comparisons of the experimental group with the two control groups; other factors have had an important effect here. The subscores for "insufficient data" and for "decrease of wrong predictions" bear out the generalization that amount of experience is important, since the experimental group had some experience with qualifying prediction during the study of interpretation of data, and the decrease in wrong predictions is a reflection of this experience. The remaining two abilities were not taught and contained few experimental elements that were taught; the reliabilities of differences are low, and tend to favor the control groups as much as the experimental.

2. *Learning of abilities: verbalization and rationalization.*—The factor here to be considered is perhaps the most important element in the learning of the abilities studied. Its importance, incidentally, has been assumed throughout the discussion in the statements as to whether a given ability was "taught". To be "taught", there must be both opportunity for experience and verbalization of that experience. The ability to interpret data is an outstanding example of an ability that was verbalized and rationalized by the students. The ability was defined as 17 specific behaviors. The students learned to recognize the types of situations in which each behavior applied, and they learned criteria or maxims or conventions by which these specific behaviors could be guided. Learned in this way, the skills became independent of any particular content, as is well shown by the fact that the content used in learning the ability was all chemical subject matter,

whereas the evaluation instrument dealt exclusively with non-chemical subject matter. This is striking evidence of skill learning divorced from content determiners. The evidence that the ability was rationalized will be presented in connection with affect as a factor in learning; in brief, the rationalization was sufficiently great to offset the effects of an attitude whose effects pervade most of the other learnings, even though this attitude may have been developed primarily in connection with interpretation of data!

3. *Learning of abilities: attitude.*—In the original report the writer has attempted to explain various features of the patterns of scores in terms of the development of an attitude of overcaution by the experimental group. The problem is complicated somewhat by the fact that the macro group developed a considerable portion of the opposite attitude, that of going beyond the data. Since these two effects operate in the same situations, it is difficult to distinguish either as a cause of differences, except when it can be measured separately. For this reason, the following discussion is confined to the comparisons between the semi-micro and experimental groups.

Scores for "overcaution" and for "beyond data" were found with regard to three different abilities. In each case, "overcaution" was measured as the number of rejections of the opportunity to come to a decision when the evidence was adequate for a decision to be made. "Beyond data" was the opposite characteristic, the number of decisions recorded for which the data were insufficient.

The initial scores for the three groups, calculated as the percent of opportunities for displaying the trait, are presented in Table V.

The groups are seen to much alike, but very different from test to test. Assuming that chance plays a large part, one might attempt to account for the differences between tests as a function of different numbers of ways available for displaying the traits on each item. Calculation shows that the expected ratios of "beyond data" to "overcaution," on the basis of chance alone, are as presented in Table VI.

It is thus demonstrated that taking account of differences in numbers of opportunities to reveal the traits (by using per cent of possible opportunity), and taking account of the number of ways of marking each item, still

TABLE V
INITIAL SCORES FOR "OVERCAUTION" AND "BEYOND DATA"

		Overcaution			Beyond Data		
		Experi- mental	Semi- micro	Macro	Experi- mental	Semi- micro	Macro
III	Interpretation of data.....	13.54	12.80	15.48	47.46	52.54	48.59
VI	Planning of Experiments..	32.49	38.02	34.68	26.22	27.26	23.40
II	Predicting Chemical Occurrences..	9.46	15.89	7.68	44.44	48.81	44.05

TABLE VI
EXPECTED RATIOS OF "BEYOND DATA" TO "OVERCAUTION," ON THE BASIS OF CHANCE

Beyond data Overcaution	Expected by chance alone	Experimental	Observed Semimicro	Macro
III	1.20	3.5	4.1	3.1
VI	1.03	0.81	0.72	0.68
II	3.00	4.7	3.1	5.7

leave a large discrepancy between relative amounts of "overcaution" and "beyond data," as measured by the three tests. The necessity for reporting the name of the test along with the description of a student's status is thus apparent.¹²

It is found that the experimental group failed to increase in overcaution on interpretation of data, as compared with the semi-micro group, that it increased very reliably in overcaution on planning of experiments, and that it indicated some increase in overcaution in prediction. This is particularly revealing, because the importance of proper qualification of data was developed mostly in connection with experiences relevant to interpretation of data. In terms of time spent, the students used probably between 5 and 10 times as much in situations conducive to learning to plan experiments as they did in situations directly relevant to interpretation of data. Little time was spent in gaining experiences relevant to prediction. The above description may be summarized as follows:

- It is believed that the students became aware of overcaution mainly with regard to interpretation of data, and that this awareness resulted in the development of an overcautious attitude in general.
- The experiences relative to interpretation of data were sharply focused and

were both rationalized and verbalized. This resulted in an increase in general accuracy without the development of overcaution on this test.

- There were more experiences relative to planning experiments than to interpreting data. "Planning experiments" is a paper-and-pencil test whose situations are very similar to the learning situations. Accuracy increased as a result of much experience incompletely rationalized. The attitude of overcaution resulted in increased errors of this sort, even though accuracy in general increased.
- There were few experiences relevant to prediction. The major shift revealed was an increase in overcaution, which is believed to be evidence that a general attitude was developed.

The conclusion is that some experience, completely rationalized, may be as effective as much experience incompletely rationalized in the development of general skill abilities. Furthermore, a general attitude can be built, but its operation may be inhibited in those situations that have been completely rationalized. Even in studied situations, if rationalization is not complete, the attitude may exert control over the behavior, and in new situations also the attitude may be much more effective than chance in determining the nature of the behavior elicited.

¹² Comparisons of the initial scores for "overcaution" and "beyond data" on these three sections reveal a much overlooked point: that attempts to describe the status of a student with regard to these two traits are meaningless except as a function of a particular test.

The general conclusion of detailed study of the learning of the five abilities¹⁸ is that there is experimental evidence for the importance of learning of the factors:

- a. Frequency of relevant experience.
- b. Intensity and multi-perception of experience.
- c. Function of learned behavior.
- d. Verbalization and rationalization of learned skills.
- e. Development and effect of attitudes.

This constitutes a check on the validity of similar methodological principles developed in the theory of science teaching.

III. SUGGESTIONS FOR TEACHING CHEMISTRY

As a result of this experiment, some proposals for the improvement of instruction in chemistry for students similar to those studied may be offered. The following suggestions receive direct support from the experiment:

1. Insofar as possible, plan activities requiring group planning and execution with responsibility divided among individuals.
2. Develop experiments in class discussion. The saving of time in the laboratory would be considerable, if twenty minutes of the quiz hour could be devoted to developing the rationale and technique of each experiment.
3. Follow up each experiment in class discussion, formulating as a class the best possible statements of conclusions, and identifying the assumptions and sources of experimental error in the experiment.
4. Widen the range of activities in the quiz-laboratory hour to provide experiences conducive to the development of a broader range of abilities.
5. Use the development of chemistry as the basis for developing understanding of scientific method, and provide real opportunity for the student to use scientific method in planning and following up experiments.
6. Eliminate or re-design the test-tube survey type of experiment, except when it can be shown that the range of superficial experiences with a wide variety of chemicals results in valid and important generalizations.
7. Attack the basic quantitative problems through supervised drill and class analysis of principles and methods. Time may be

found for this by changing the quiz-reviews to short diagnostic tests and discussions of the test problems.

8. Teach a verbalized rationale along with all skills. The rationale should include criteria for deciding what behaviors are appropriate to solve the problem in accord with desirable values.

The following suggestions are reasonable extensions of the principles demonstrated in this experiment:

9. Theory needs to be re-considered as an objective for the course. The learning of the theory is incommensurate with the time and effort spent in trying to teach it. It might be better either to omit all theory not needed by the student or to redesign activities, so that the role of theory is more explicit. More emphasis upon explanation and prediction would tend to have this latter effect.

10. Consider the possibility of replacing the lectures with demonstration-discussions or of eliminating them completely. This consideration should embrace:

- a. The amount of the lecture time devoted to outlining theory that is not learned.
- b. The amount of the lecture time devoted to performing experiments that either are or could be performed with more profit by the students.
- c. The amount of the lecture time devoted to attempts to build, through verbal instruction, abilities whose development depends upon participation in specially planned activities.
- d. The amount of lecture time devoted to outlining the material in the book. (Perhaps having the student outline the book during supervised study periods would be more profitable to him).

11. Plan each experiment (except those whose function is exploratory) to involve prediction of results and comparison of results with prediction.

12. Teach fewer and larger generalizations through providing opportunities for experience with physical and biological material as well as chemical.

13. Give diagnostic pre-tests and plan sufficiently "rich" activities that individual students can be guided into having experiences most profitable to them.

¹⁸ Presented in Chapters VI-X of the complete report.

14. Use the book only in connection with definite assignments to be fulfilled through and at the time of reading the book. Outlining, answering guide questions, using data in the book to support arguments, and writing out definitions of words may be suggested to guide the reading of the book. A reading and problem clinic might well be established to provide assistance for the students needing it. Since much of the material presented in the text might well be worked out under supervision by the student, it might be well to eliminate the book and replace it by a handbook or a shelf of reference books.

The following suggestions conform with the theory of science teaching, but go completely beyond present practice:

15. Introduce into the course much more information about how plants operate, their relationships with special interest groups and classes in the population, and their societal privileges and responsibilities. Also help the student to understand what the job of being a chemist is like.

16. Plan experiments to reveal the functions of chemistry in industry (prevention of corrosion, synthesis of compounds for specific purposes, development of substitutes, action of fertilizer, analysis of food products and other consumer goods, and the like). These topics should then be placed in their societal context through outside reading, lecturing, community survey, analysis of advertising, or other means.

17. Teach the dignity and intelligence of man by showing that scientific principles are made rather than "discovered" by man.

18. Plan a long-range program of evaluation-guided research to develop the most effective possible teaching procedures.

SUMMARY

In the experiment one general type of outcome, the ability to think scientifically, has been studied through comparisons of two instructional methods. The control method was believed to be fairly representative of the most usual instruction in freshman college chemistry. The experimental method

was a modification of the control methods and was guided insofar as possible by a reasonably comprehensive stated theory of science education. The experimental method as here applied was delimited by the need for control in the comparisons, by the present stated objectives of the control method, and by limits in present understanding of teaching techniques. It is believed to have represented a desirable step, but to have nowhere approached the stated theoretical possibilities in science education.

The evidence in the experiment was obtained from a four-hour test battery comprising six sections. Each section presumably elicited specific behaviors believed to be pertinent to a general ability. A rationale for each ability and a description of the evaluation instrument used in its appraisal have been presented. The description of achievement of each ability was obtained from a pattern of scores; there were 31 scores obtained from the test battery.

The major findings of the experiment were based upon the agreement between 31 specific hypotheses and observed scores obtained by students under the two methods of instruction. Since the students were tested at the middle of year, as well as at the beginning and end, it has been possible in some cases to describe what appears to be the course of learning of certain behaviors. Interpretations as to the validity of certain facets of the theory resulted from comparisons by the method of analysis of covariance of achievement in the various abilities under the two methods. It is believed that the theory has been checked in certain important aspects dealing with the role of experience and the value of rationalization of behavior.

Further interpretations of the educational significance of the differences in outcomes under the two methods have been discussed, the general nature of some desirable further research has been considered (in the original report), and a series of suggestions has been offered for improvement of instruction in freshman chemistry within the existing administrative framework.

